CogView: Mastering Text-to-Image Generation via Transformers

Year: 2021 | Citations: 905 | Authors: Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou

---

Abstract

Text-to-Image generation in the general domain has long been an open problem, which requires both a powerful generative model and cross-modal understanding. We propose CogView, a 4-billion-parameter Transformer with VQ-VAE tokenizer to advance this problem. We also demonstrate the finetuning strategies for various downstream tasks, e.g. style learning, super-resolution, text-image ranking and fashion design, and methods to stabilize pretraining, e.g. eliminating NaN losses. CogView achieves the state-of-the-art FID on the blurred MS COCO dataset, outperforming previous GAN-based models and a recent similar work DALL-E.